THE USE OF LOGISTIC REGRESSION IN DIAGNOSTIC AND PROGNOSTIC PREDICTION IN A MEDICAL INTENSIVE CARE UNIT

Erik R. Skinner, G. Octo Barnett, Daniel E. Singer,
Albert G. Mulley, Robert A. Lew, Susan A. Stickler,
Mary M. Morgan, George E. Thibault

The Laboratory of Computer Science and the Medical Practices Evaluation Unit, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114

A stepwise logistic regression model was used to predict the one year survival of patients admitted to a medical intensive care unit. Two methods of validation were used to test for stability, overtraining, and the effects of additional variables: cross validation using separate training and validation sets and the jackknife technique. The effect of correlation between variables and relative frequencies in the jackknife subgroups on the model is discussed. The use of various cut-off values to change the sensitivity and specificity of the model is examined.

INTRODUCTION

A number of statistical techniques have been utilized in analyses from clinical databanks to provide diagnostic or clinical predictions. Such studies done by the Laboratory of Computer Science of the Department of Medicine at the Massachusetts General Hospital have included several methods, among them sequential Bayes[1], Bahadur's technique[2], multiple regression, discriminant analysis and logistic regression. This paper discusses our experience in using the logistic regression model for prediction of survival at one year following discharge, for a specific cohort of patients admitted to an intensive care unit.

METHODS

Logistic Model: The objective of logistic regression is to produce a formula for predicting the probability of success for a given outcome for each patient in a study. One could predict this probability to be simply the rate of the outcome in the study group. Logistic regression attempts to improve on this prior probability by adjusting the predicted probability for a given patient according to specific covariate values. (e.g., sex = male, age = 49, CPK = 50).

Letting X denote the vector of covariates and β the associated coefficients, the logistic model relates the individual's probability p to his or her covariate values as follows:

$$\text{Log}\left(\frac{p}{1-p}\right) = \beta_0 x_0 + \beta_1 x_1 \ldots + \beta_K x_K = \beta' x \qquad [1]$$

β_0 is the intercept coefficient and x_0 is identically equal to 1 for all records. Inverting this equation one obtains

$$p = \exp(\beta' X) \;/\; (1+\exp(\beta' X)) \qquad [2]$$

The probability of the data or the likelihood function of the vector β, $L(\beta)$, is essentially a binomial likelihood function, with the probability of success for each record modified by its covariate values. Let the sample contain n persons. Then precisely

$$L(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i} \left(1-p(x_i)\right)^{1-y_i} \qquad [3]$$

Substituting equation [2] into $L(\beta)$ where $p(x_i)$ corresponds to p for record i with covariates $x=x_i$ we obtain

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp\left(\left(\beta' x_i\right) y_i\right)}{1+\exp(\beta' x_i)} \qquad [4]$$

The estimates of β are then obtained by maximizing equation [4] as a function of β using the Newton-Raphson method[3]. The resulting estimates $\hat{\beta}$ are the maximum likelihood estimates.

The coefficients generated by the logistic regression can be applied to a population resulting in a logit value for each person in the population. Since the range of this logit value is from 0 to 1, the logistic regression is easily applied to a problem with a dichotomous outcome.

The logistic regression used in this analysis is written in MUMPS to run on a PDP-15. Variables are added to the model using a stepwise procedure with variable selection based on the gain obtained by computing the increase in log-likelihood attributed to the variable. The level of significance for the gain is measured by the likelihood ratio test[4]. The maximum number of variables allowed in a single regression run is 99, although response time makes such a run impractical. A significant improvement in response time can be achieved by using dichotomous variables.

0195-4210/80/0000-0222$00.75 © 1980 IEEE

Clinical Database: Data has been collected on all patients admitted to an 18-bed intensive care unit (ICU) at the Massachusetts General Hospital[5,6]. Data collection began on July 18, 1977 and to date over 4,000 records have been entered into the database, with more than 400 data items entered into each record. Included in the data are demographic information, medical history, physical findings and follow-up information.

In addition, extensive clinical data were collected on all cardiac patients admitted to the ICU for one year beginning January 29, 1978. A subset of patients from this one-year group was identified as having been admitted to the unit with a diagnosis of either acute myocardial infarction, suspected myocardial infarction, coronary insufficiency or unstable angina. Of these patients 563 were discharged alive from the hospital and were followed-up within one year. From this group, 417 patients with chest pain on admission were selected for further consideration. A final subset of 397 patients was derived by eliminating those patients without values for CPK, LDH, or SGOT and those who were not given an EKG on admission. This final subset became our study group for examination of the use of stepwise logistic regression to build a model for prediction of survival. Of the 397 patients, 46 were deceased at follow-up; 351 were still alive.

Variable Selection: Preprocessing of the data was necessary to reduce the variables to a workable number. A chi-square test was used on all categorical variables and a t-test was used on all continuous variables to determine the best candidates for predicting survival. All variables were then converted to dichotomous variables. Cut-off points used to convert some continuous variables into dichotomous variables were determined using maximum Kolmogorov-Smirnov test[7]. Other continuous variables, such as CPK, were split into several variables, each having a range. (E.g., CPK< 50, CPK>50 and < 100, CPK >100 and < 250, CPK >250.) At the end of this process, there remained 296 dichotomous variables eligible for use in the model.

These 296 variables were grouped into systems such as medical history, physical exam, etc., each of which contained from 18 to 24 variables. The grouping by systems was used in the final analysis to enhance recognition of variables which were highly correlated. Logistic regression was run on each system to determine the best variables for the final analysis.

Validation of the Model: Two approaches were used to validate the model. The first approach involved dividing the total set of patients into two subsets, a training set on which the model was built and a validation set to test the accuracy of the model. The second approach involved dividing the total set into subsets and using the jackknife technique[8,9] to check the stability of the model.

The performance of the stepwise logistic regression model was measured by comparing the probability of death predicted by the model against the actual outcome. To do this the predicted probabilities were converted to a dichotomous form and the subsequent comparison made via a 2 x 2 table. For example, if the cut-off value is 0.2, any patient with a predicted probability greater than 0.2 was considered to have a positive outcome; any patient with a predicted probability less than or equal to 0.2 was considered to have a negative outcome. Given a specific cut-off value, a 2 x 2 table (see Table 1) was generated and by sweeping through several cut-offs between 0 and 1, a series of tables was generated. (The predictive probabilities generated using the logistic model fall in the range of 0 to 1.)

TABLE 1
2X2 TABLE

	OUTCOME +	OUTCOME -	
TEST +	a	c	a+c
TEST -	b	d	b+d
	a+b	c+d	

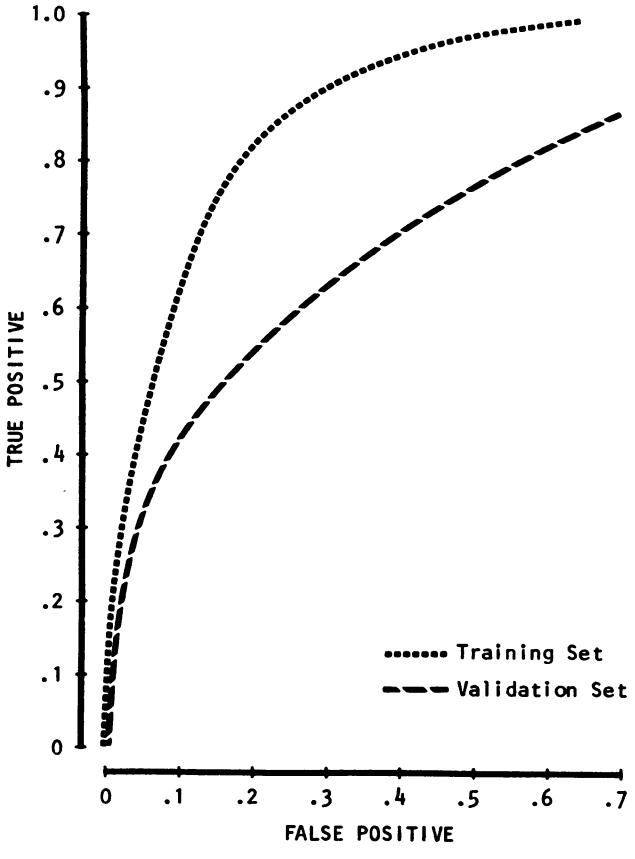
Taking the proportion of true positives (a/a+b) and the proportion of false positives (c/c+d) derived at several cut-off values, an R.O.C. curve[10,11] was generated. Comparison of R.O.C. curves generated by a given model on the training and validation sets provides a measure of model performance. (See Figure 1)

The clinician may select the utilities best suited for his or her own application. A low cut-off value yields a high true positive but also a high false positive rate, while a high cut-off value yields a low true positive and low false positive rate. Once the clinician determines suitable values for true positive ratio and false positive ratio, an appropriate cut-off value for the model can be determined. The model's performance at that cut-off is shown by the R.O.C. curve generated by the model.

Additional Variables: The logistic regression employed was a stepwise procedure which added a new variable whenever a significant gain resulted. Changes in sensitivity and specificity provided a measure of the usefulness of adding variables to the model. Holding the specificity constant, the sensitivity of the model was recorded after each variable was added. Likewise, the specificity of each model at a constant sensitivity was recorded. The point at which the sensitivity and/or specificity decreased demonstrated when additional variables showed little new information. Careful attention was paid to the significance of the increases in sensitivity and specificity. The effect of additional variables was examined in the model generated by the training set on both the

FIGURE 1

ROC CURVE



training set and validation set populations and also in the model generated by the total set using the total population.

Jackknife Technique: The jackknife procedure developed by Quenoille[8] and extended by Tukey[9] requires dividing the total set into several subsets, each having approximately equal numbers of patients and approximately equal prior odds. The jackknife population groups were formed by taking the total set of patients minus one of the subsets defined above. One model using the total set and one model for each of the jackknife groups were generated. The jackknife models and the total set model were compared using the information content as described by Shannon[12]. The information content was measured at a cut-off point common to all the models; then Tukey's jackknife formula[13] was used to compute a best estimate for the total model.

RESULTS

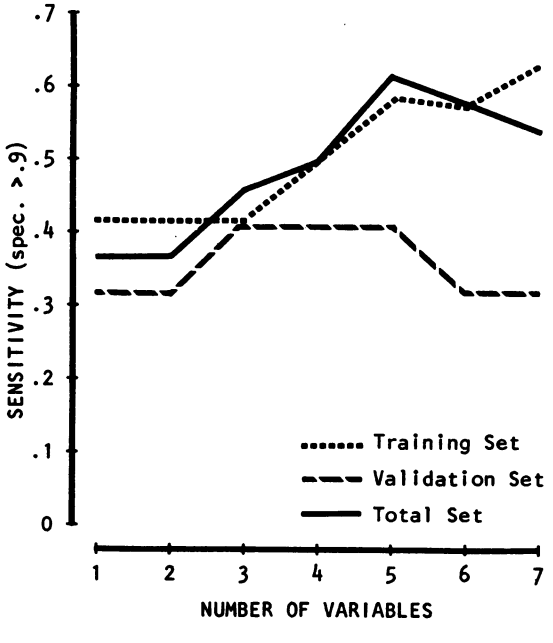
Variable Selection: The logistic regression runs made to determine the variables used for the final model yielded 24 variables which had a gain of greater than 0.82 (gain = 0.82 corresponds to a $p < 0.20$ significance level as determined by the likelihood ratio test[4]). The total set model and the training set model were created using these 24 variables. However, to reduce the amount of execution time, the jackknife models were created using only those variables with a gain greater than 0.82 in the model generated by the total set.

Cross Validation: To establish the training and validation sets for the cross validation procedure, each of the 397 patients in the total data set used was randomly assigned to either the training set or the validation set. The result was a training set of 204 patients, 24 of whom were deceased at the time of follow-up, and a validation set of 193 patients, 22 of whom died within one year. The stepwise logistic regression was run and a model generated using 5 variables, each gaving a gain greater than 1.92. (A gain of 1.92 corresponds to a significance level of $p < .05$ as determined by the likelihood ratio test).

The stepwise logistic regression generated a model finely tuned to the relative frequencies of the covariates in the training set. The R.O.C. curve demonstrated the more conservative performance of the validation set when compared with the training set. When this overtraining occurred the performance of the validation set was used as a conservative measure of the model's performance.

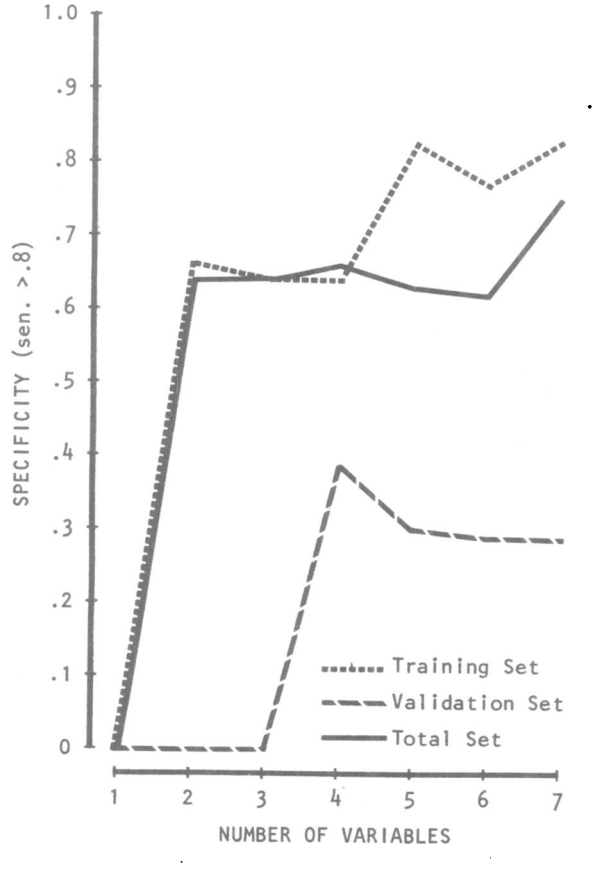
Additional Variables: Figure 2 demonstrates the contribution additional values made to the model, with respect to sensitivity.

FIGURE 2



The graph shows the highest sensitivity achieved if the specificity was greater than .9 for the seven models (one for each variable added to the model) generated for each of the three populations: the training set, validation set and a total set made up of all 397 patients. After the fifth variable had been added, the models applied to the training set showed no increase in sensitivity (the gain achieved by adding the seventh variable was insignificant, hence the rise in sensitivity at the seventh variable was attributed to noise). The models applied to the validation set showed a leveling at the third variable and a decrease when the fifth variable was added. Figure 3 is a graph of the specificity given a sensitivity of .8 or greater for the seven models for each population.

FIGURE 3



The models applied to the training set demonstrated no new increase in specificity after the fifth variable (again, the seventh variable introduced noise), while the graph of the validation set showed no change in specificity after the fourth variable. The graphs of sensitivity and specificity for the models generated using all 397 patients had results similar to the training set models.

The addition of a sixth variable gave no new information to the model generated by the training set. The decision of how many variables are to be included in the model depends upon the outcome to which the model is applied. The user must decide within which range of sensitivity and/or specificity he/she wishes to operate, and make the appropriate model choice. A lower acceptance criterion for specificity or sensitivity may allow more variables to be added.

Jackknife Technique: For the jackknife, the 397 patients in the total set were divided into six subsets, each having approximately 66 patients and each having approximately equal prior odds of death within one year. Six different jackknife groups were formed by combining a unique group of five of the six subsets defined above. Thus, each jackknife group contained 5/6 of the total set (331 ± 1). The mortality rate in these groups was 11.6 ± 0.15%. A model was generated using the total population and 24 variables in a stepwise logistic regression. Allowing this model to expand as long as added variables produced gains greater than .23 (gain of .23 corresponds to $p < .5$) resulted in a ten-variable model. These ten variables were then used for the jackknife analysis.

A total of seven models was generated for the jackknife procedure: six models, each one using one of the jackknife groups as a population, and one model using the total set of 397 patients. Applying the six jackknife models to their respective populations, the information content[12] for each model was generated using a cut-off point equal to the prior odds of death in one year (.116). The same measurement was done on the total set using the model generated by the total set, resulting in an information content of .1147. Applying the information content of Tukey's jackknife formula[13], a best estimate of .1054 with a 95% confidence interval of -.0187 to .2294 was achieved. A sequential Bayes model[1] applied to the same data set resulted in a lower best estimate for information content but a narrower 95% confidence interval.

The six models generated by applying the stepwise logistic regression to the six jackknife population groups differed with respect to the number of variables selected, the identity of the variables selected and the order of their selection (Table II). This variability resulted from changes in the relative frequencies of the variables within each jackknife group and the correlation of variables with one another.

TABLE II

ORDER OF VARIABLE SELECTION

VARIABLE	JACKKNIFE GROUP 1	2	3	4	5	6	TOTAL SET
1	-	-	1	-	-	1	1
2	5	2	2	-	2	2	2
3	2	1	-	2	1	3	3
4	4	3	3	-	5	5	4
5	3	5	5	3	4	4	5
6	1	4	-	1	6	6	6
7	6	-	4	-	7	-	7
8	-	-	-	-	3	-	-
9	-	-	6	-	8	-	-
10	-	-	-	-	-	-	-

- denotes not selected for model

The information content of variable 6, for example, was .0367, in jackknife group 2 and .0744 in jackknife group 4. Similarly, the sensitivity of variable 6 changed from .6154 to .7105 in the respective jackknife groups and the specificity changed from .7235 to .7705. This variation in information content, sensitivity and specificity occurred in all variables throughout all jackknife groups.

The effect of variable correlation became more complex as more variables were added to a given model. The first variable chosen for a model was selected on its ability to discriminate between the positive outcome (death at follow-up) and the negative outcome (survival at follow-up). When the stepwise logistic regression evaluated the remaining variables for possible selection as the second variable in the model, the correlation between the variable already chosen and the potential variable was taken into account by the logistic regression algorithm.

If two variables were strongly correlated with one another and one of the variables was also chosen by the regression for the model, the second variable would be an unlikely candidate for addition to the model. The first variable selected would account for the power of the second, hence little new information could be gained by adding the second variable. For example, at the first step of the regression for the model generated using the jackknife group 1, variable 6 (history of congestive heart failure) was ranked first with a gain of 13.51, while variable 1 (pulmonary edema at admission) with a gain of 11.14 was ranked second and variable 3 (wet rales (not basilar)) was ranked third with a gain of 8.34. Variable 6 was selected as the first variable in the model and the remaining nine variables were tested with the new one-variable model for selection of a second variable.

In the second step, variable 3 was ranked first (gain = 5.26) with variable 1 ranked second (gain = 4.67). The shift in rank of variables 1 and 3 resulted from the correlation of variables 6 and 1. (Of the 35 patients in jackknife group 1 with variable 1 positive, 27 (77%) patients also had variable 6 positive, while 12 of the 19 (63%) patients with variable 3 positive also had variable 6 positive). While both variables 1 and 3 were correlated with variable 6, variable 1 was affected more by the correlation between variables.

Once two variables have been selected for the model, the correlation problem becomes more complicated. The logistic regression had to take into account the correlation between the candidate variables and the variables already chosen. In the example cited above, variable 1 had a gain of 2.01 on the third step of the regression and was ranked fifth out of the eight remaining variables. As the model generation progressed, variable 1 was never chosen.

DISCUSSION

Statistical models for clinical prediction must be judged primarily on their ability to accurately predict outcomes for new patients. The ability of the logistic regression model discussed in this paper to predict mortality has been demonstrated by its information content and R.O.C. curve. However, the model clearly has drawbacks. In particular, overtraining was evident as model performance decreased with the validation set of patients and the information content of the individual jackknife group models varied widely. In addition, correlation among variables led to instability in the choice of predictors for the different jackknife groups. Nonetheless, the models retained predictive power.

Clinical reality may not conform to the assumptions of the logistic regression technique. For example, variables may have multiplicative effects not accounted for by the simple linear model. Moreover the logit form may not approximate the true relationship between the probability of a given outcome and the values of a predictor. In every case the nature of the clinical problem and the variables involved must be carefully assessed. Defining new variables and combining other statistical techniques with the logistic regression may help solve these problems.

References

1. Morgan MM, Barnett GO, Skinner ER, Lew R, Mulley AG, Thibault GE. The use of a sequential Bayesian model in diagnostic and prognostic prediction in a medical intensive care unit. These proceedings.

2. Spitzer PE. Non-parametric computer-aided diagnosis: Bahadur's technique. M.S. thesis, Mass. Institute of Technology May 1980.

3. Pennington RH. Introductory Computer Methods and Numerical Analysis. Macmillan Co, New York 1965.

4. Cox DR. Analysis of Binary Data. Chapman and Hall, London 1970.

5. Thibault GE, Mulley AG, Barnett GO, Goldstein RL, Reder VA, Sherman EL, Skinner ER. Medical intensive care: Patients, interventions, costs and outcomes. N Engl J Med. 1980, 302:938-942.

6. Mulley AG, Thibault GE, Hughes RA, Barnett GO, Reder VA, Sherman EL. The course of patients with suspected myocardial infarction. N Engl J Med. 1980, 305:943-948.

7. Siegal S. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill Book Co., New York 1956.

8. Quenoille MH. Appropriate tests of correlation in time series. J R Statist Soc. 1949, B 11, 68-84.

9. Tukey JW. Bias and confidence in not-quite large samples. Ann Math Statist. 1958, 29:614. (Abstract)

10. McNeil BJ, Keller E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med. 1975, 293:211-215.

11. Metz CE, Goodenough DJ, Rossman K. Evaluation of receiver-operating characteristic curve data in terms of information theory, with applications in radiography. Radiography. 1973, 109:297-303.

12. Shannon EC, Weaver W. The Mathematical Theory of Communication. University of Illinois Press. Urbana, Illinois, 1949.

13. Mosteller F, Tukey JW. Data Analysis and Regression: Second Course in Statistics. Addison-Wesley Publishing Co., Reading, Massachusetts, 1977.